D90–D93 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 23 October 2012*
doi:10.1093/nar/gks990

DoriC 5.0: an updated database of *oriC* regions in both bacterial and archaeal genomes

Feng Gao*, Hao Luo and Chun-Ting Zhang*

Department of Physics, Tianjin University, Tianjin 300072, China

Received August 14, 2012; Revised September 27, 2012; Accepted September 29, 2012

ABSTRACT

Replication of chromosomes is one of the central events in the cell cycle. Chromosome replication begins at specific sites, called origins of replication (*oriCs*), for all three domains of life. However, the origins of replication still remain unknown in a considerably large number of bacterial and archaeal genomes completely sequenced so far. The availability of increasing complete bacterial and archaeal genomes has created challenges and opportunities for identification of their *oriCs in silico*, as well as *in vivo*. Based on the *Z*-curve theory, we have developed a web-based system Ori-Finder to predict *oriCs* in bacterial genomes with high accuracy and reliability by taking advantage of comparative genomics, and the predicted *oriC* regions have been organized into an online database DoriC, which is publicly available at http://tubic.tju.edu.cn/doric/ since 2007. Five years after we constructed DoriC, the database has significant advances over the number of bacterial genomes, increasing about 4-fold. Additionally, *oriC* regions in archaeal genomes identified by *in vivo* experiments, as well as *in silico* analyses, have also been added to the database. Consequently, the latest release of DoriC contains *oriCs* for >1500 bacterial genomes and 81 archaeal genomes, respectively.

INTRODUCTION

The identification of replication origins will be helpful to reveal the regulatory mechanisms of the initiation step in DNA replication (1,2) and discover new broad-spectrum antibacterial drugs (3). Based on the *Z*-curve theory (4), we have developed a web-based system Ori-Finder for finding *oriCs* in bacterial genomes with high accuracy and reliability (5), and the predicted *oriC* regions in bacterial genomes have been organized into an online database DoriC (6). Based on the database, putative origins of replication in *Sorangium cellulosum*, *Microcystis aeruginosa* (7) and *Cyanothece* 51142 (8), which could not be determined by using standard GC skew, have been identified by taking advantage of comparative genomics. The application of the proposed *oriC* selection criteria and the comparison of different cyanobacterial strains may also gain insight into the replication origins in other cyanobacteria (9). As the database was constructed in 2007, we noticed that the replication origins of *Anabaena* sp. PCC 7120 (10), *Cytophaga hutchinsonii* ATCC 33406 (11) and *Synechococcus elongatus* PCC 7942 (12) have been confirmed by experiments, which are all consistent with our predictions in DoriC. Because of continuous updates, our database has been widely used in the comparative genomics analysis. For example, as a source of data, DoriC has been used in the study of the relationship between the functionality of essential genes and gene strand bias in bacterial genomes (13), in the analysis of nucleotide compositional asymmetry between the leading and lagging strands of bacterial genomes (14), in the investigation of the association between growth-related traits and minimal generation times (15), in an algorithm for prediction of putative essential and core-essential genes in *Mycoplasma* genomes (16), in the research on coordination of spatiotemporal gene expression during the bacterial growth cycle (17) and in the study of the variation in terms of the percentage of leading strand genes across different bacteria (18), etc. It is expected that the new release of the database, DoriC 5.0, will promote the study of *oriCs* in both bacteria and archaea.

DATABASE UPDATES

In the current release, the database has been significantly improved compared with the initial release, and the main advances include (i) inclusion of *oriCs* in more bacterial genomes that increased from 435 to 1528; (ii) inclusion of *oriCs* in 81 archaeal genomes; (iii) inclusion of detailed information about repeats in *oriCs* identified by REPuter program (19); and (iv) addition of URLs that

*To whom correspondence should be addressed. Tel: +86 22 2740 2987; Fax: +86 22 2740 2697; Email: fgao@tju.edu.cn
Correspondence may also be addressed to Chun-Ting Zhang. Tel: +86 22 2740 2987; Fax: +86 22 2740 2697; Email: ctzhang@tju.edu.cn

link to NCBI Map Viewer (20) or UCSC Archaeal Genome Browser (21), which are useful to explore and discover the conserved features around the *oriC* region. Consequently, the latest release of DoriC contains *oriCs* for >1500 bacterial genomes and 81 archaeal genomes, which can be accessed from http://tubic.tju.edu.cn/doric/.

DATABASE DESCRIPTION

Replication origins in bacteria

To identify *oriC* regions of unannotated bacterial genomes, we have developed a web-based system, Ori-Finder, based on an integrated method comprising gene identification, analysis of base composition asymmetry using the *Z*-curve method, distribution of DnaA boxes, occurrence of genes frequently close to *oriCs* and phylogenetic relationships. Consequently, the predicted *oriC* regions have been organized into an online database, DoriC. Based on DoriC, the relationships between the conserved features associated with the *oriC* regions, such as adjacent genes, DnaA boxes, etc., and the taxonomic levels of the corresponding bacteria have been summarized. For example, detailed analyses have shown that the consensus sequence of the DnaA boxes in *oriC* regions and the distribution of genes around *oriCs* are strongly conserved among the bacteria in the phylum cyanobacteria (7,8). The feature that the *oriC* is adjacent to *dnaN* gene, which encodes the beta clamp processivity factor, has been found to be universal among the bacteria within the phylum cyanobacteria, and the 'species-specific' DnaA box motif for the phylum cyanobacteria is 'TTTTC CACA' instead of 'TTATCCACA', the DnaA box motif of *Escherichia coli*. These strongly conserved features indicate that the *in silico* identified *oriCs* are reliable, as they have been confirmed by comparative genomics approaches. This observation also shows that if the *oriC* for one of the bacteria in the phylum cyanobacteria is confirmed experimentally, the *oriCs* for the other bacterial genomes in this phylum may be confirmed simultaneously. As we expected, the experimentally confirmed replication origins of *Anabaena* sp. PCC 7120 (10) and *S. elongatus* PCC 7942 (12) in the phylum cyanobacteria are all adjacent to the *dnaN* gene, which encodes the beta clamp processivity factor. Therefore, the proposed rules may be helpful to predict the *oriC* regions for some bacteria without complete genomes in the phylum cyanobacteria. In addition, the application of the proposed rules derived from DoriC would speedup the experimental confirmation and functional analysis of *oriCs* in bacterial genomes. Because of the rapid growth in the number of sequenced bacterial genomes, the replication origins for those unsubmitted to GenBank or not deposited in DoriC temporarily can be predicted by Ori-Finder firstly, which now has been used to analyze ~30 newly sequenced bacterial genomes.

Replication origins in archaea

The *Z*-curve analysis has been used to identify one replication origin in the genomes of *Methanocaldococcus jannaschii* (22) and *Methanosarcina mazei* (23), two replication origins in the *Halobacterium* species NRC-1 genome (24), which have been confirmed by *in vivo* experiments (25,26) and three replication origins in the *Sulfolobus solfataricus* P2 genome (24), which have been later confirmed experimentally (27,28). Here, we collected the information of *oriCs* provided in the literature, such as the *oriC* sequences, origin recognition boxes (ORB) motifs, uncharacterized motif sequences, etc., which were identified by *in vivo* experiments (25–34), as well as *in silico* analysis (4,22–24,35). In addition, we also predicted some new replication origins by *Z*-curve method, with the aid of homologous sequence search against the known replication origins, analysis of ORB motifs and repeats, *cdc6* gene location, etc. Consequently, *oriC* regions in 81 archaeal genomes identified by *in vivo* experiments, as well as *in silico* analyses, have been added to our database. The number of *oriCs* in archaea is correlated with the phylogeny, which has been summarized in detail in the 'Introduction' section of the (34). Based on our results in DoriC, it shows that there is one replication origin in the genomes within the order *Methanococcales* (11 genomes) and within the class *Thermococci* (12 genomes), and three replication origins in *Sulfolobus* species (13 genomes). Our results and the *Z*-curves also show that the archaea within the Crenarchaeota phylum contain multiple origins, although some origins could not be determined at the sequence level currently. For example, *Pyrobaculum calidifontis* has been experimentally characterized to contain four replication origins, which is the highest number detected in a prokaryotic organism (34). However, only one origin can be determined at the sequence level (34). During the course of the prediction, we found that the location of some putative replication initiator gene besides *cdc6* gene can be helpful to the *oriC* prediction in some cases. For example, in the genome of *M. jannaschii*, an ORF (MJ0774), annotated as a hypothetical protein, is a distant homolog of the Cdc6 protein in fact (22). The name Mc-pRIP for the putative replication initiator protein in *Methanococcales* has been used here for MJ0774 and related proteins to distinguish it from bona fide orthologous *Cdc6*. We also found the genes, which encode Mc-pRIP in other 10 genomes within the order *Methanococcales* (*Methanococcus aeolicus* Nankai-3, *Methanocaldococcus fervens* AG86, *Methanococcus maripaludis* C5, *M. maripaludis* C6, *M. maripaludis* C7, *M. maripaludis* S2, *M. maripaludis* X1, *Methanococcus vannielii* SB, *Methanococcus voltae* A3 and *Methanocaldococcus vulcanius* M7), were annotated as 'LysR family protein', 'regulatory protein ArsR', 'MarR family transcriptional regulator', etc. Based on the locations of these genes, the *oriCs* in the aforementioned genomes were predicted reliably, which contains almost all the features of known replication origins in archaeal genomes. URLs that link to NCBI Map Viewer or UCSC Archaeal Genome Browser (if available) are also provided, which will be useful to explore and discover the conserved features around the *oriC* region. With the availability of an increasing number of archaeal genomes, the prediction will be more accurate and reliable, as the ORB elements or genes frequently close to *oriCs* can also be

analyzed by comparative genomics, and new rules for replication origins in archaeal genomes will also be extracted in the future with the continuous update of DoriC. Here, motif-based sequence analysis tools, the multiple EM for motif elicitation (MEME) Suite (36), have been used to discover motifs in the replication origins of closely related species, e.g. the archaea from the order *Thermococcales*. Consequently, ORB motifs and some new uncharacterized motif sequences have been found by the MEME Suite and are also included in the database.

CONCLUSION

With the increased availability of completely sequenced bacterial and archaeal genomes and experimental evidence, the database will become more useful because of including more information. The application of the rules from the database will be helpful to develop new prediction algorithms of replication origins and speedup the experimental confirmation and functional analysis of *oriCs* in bacterial or archaeal genomes. Systematic and functional analysis of *oriC* regions in bacteria and archaeal genomes will also be useful for the construction of the minimum genome and regulation of growth rate and generation time of bacteria and archaea, which play a key role in the emerging field of synthetic biology. DoriC will be updated periodically to include more entries, and to integrate more information for each entry. We also welcome any feedback or corrections to help us improve the database.

ACKNOWLEDGEMENTS

The authors thank Dr Kurtz for providing the REPuter binaries. They also thank Dr Ren Zhang for critical revision of manuscript. Helpful discussions with Yan Lin and Chong Peng are also gratefully acknowledged.

FUNDING

The National Natural Science Foundation of China [31171238, 30800642 and 10747150]. Funding for open access charge: The National Natural Science Foundation of China [31171238].

Conflict of interest statement. None declared.

REFERENCES

1. Mott,M.L. and Berger,J.M. (2007) DNA replication initiation: mechanisms and regulation in bacteria. *Nat. Rev. Microbiol.*, **5**, 343–354.
2. Katayama,T., Ozaki,S., Keyamura,K. and Fujimitsu,K. (2010) Regulation of the replication cycle: conserved and diverse regulatory systems for DnaA and *oriC*. *Nat. Rev. Microbiol.*, **8**, 163–170.
3. Robinson,A., J Causer,R. and E Dixon,N. (2012) Architecture and conservation of the bacterial DNA replication machinery, an underexploited drug target. *Curr. Drug Targets*, **13**, 352–372.
4. Zhang,R. and Zhang,C.T. (2005) Identification of replication origins in archaeal genomes based on the Z-curve method. *Archaea*, **1**, 335–346.
5. Gao,F. and Zhang,C.T. (2008) Ori-Finder: a web-based system for finding *oriCs* in unannotated bacterial genomes. *BMC Bioinformatics*, **9**, 79.
6. Gao,F. and Zhang,C.T. (2007) DoriC: a database of *oriC* regions in bacterial genomes. *Bioinformatics*, **23**, 1866–1867.
7. Gao,F. and Zhang,C.T. (2008) Origins of replication in *Sorangium cellulosum* and *Microcystis aeruginosa*. *DNA Res.*, **15**, 169–171.
8. Gao,F. and Zhang,C.T. (2008) Origins of replication in *Cyanothece* 51142. *Proc. Natl Acad. Sci. USA*, **105**, E125; author reply E126–E127.
9. Welsh,E.A., Liberton,M., Stöckel,J. and Pakrasi,H.B. (2008) Reply to Zhang et al.: identification of origins of replication in the *Cyanothece* 51142 genome. *Proc. Natl Acad. Sci. USA*, **105**, E126–E127.
10. Zhou,Y., Chen,W.L., Wang,L. and Zhang,C.C. (2011) Identification of the *oriC* region and its influence on heterocyst development in the filamentous cyanobacterium *Anabaena* sp. *strain PCC 7120*. *Microbiology*, **157**, 1910–1919.
11. Xu,Y., Ji,X., Chen,N., Li,P., Liu,W. and Lu,X. (2012) Development of replicative *oriC* plasmids and their versatile use in genetic manipulation of *Cytophaga hutchinsonii*. *Appl. Microbiol. Biotechnol.*, **93**, 697–705.
12. Watanabe,S., Ohbayashi,R., Shiwa,Y., Noda,A., Kanesaki,Y., Chibazakura,T. and Yoshikawa,H. (2012) Light-dependent and asynchronous replication of cyanobacterial multi-copy chromosomes. *Mol. Microbiol.*, **83**, 856–865.
13. Lin,Y., Gao,F. and Zhang,C.T. (2010) Functionality of essential genes drives gene strand-bias in bacterial genomes. *Biochem. Biophys. Res. Commun.*, **396**, 472–476.
14. Qu,H., Wu,H., Zhang,T., Zhang,Z., Hu,S. and Yu,J. (2010) Nucleotide compositional asymmetry between the leading and lagging strands of eubacterial genomes. *Res. Microbiol.*, **161**, 838–846.
15. Vieira-Silva,S. and Rocha,E.P. (2010) The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.*, **6**, e1000808.
16. Lin,Y. and Zhang,R.R. (2011) Putative essential and core-essential genes in *Mycoplasma* genomes. *Sci. Rep.*, **1**, 53.
17. Sobetzko,P., Travers,A. and Muskhelishvili,G. (2012) Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. *Proc. Natl Acad. Sci. USA*, **109**, E42–E50.
18. Mao,X., Zhang,H., Yin,Y. and Xu,Y. (2012) The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. *Nucleic Acids Res.*, **40**, 8210–8218.
19. Kurtz,S., Choudhuri,J.V., Ohlebusch,E., Schleiermacher,C., Stoye,J. and Giegerich,R. (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.*, **29**, 4633–4642.
20. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2012) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **40**, D13–D25.
21. Chan,P.P., Holmes,A.D., Smith,A.M., Tran,D. and Lowe,T.M. (2012) The UCSC archaeal genome browser: 2012 update. *Nucleic Acids Res.*, **40**, D646–D652.
22. Zhang,R. and Zhang,C.T. (2004) Identification of replication origins in the genome of the methanogenic archaeon, *Methanocaldococcus jannaschii*. *Extremophiles*, **8**, 253–258.
23. Zhang,R. and Zhang,C.T. (2002) Single replication origin of the archaeon *Methanosarcina mazei* revealed by the Z curve method. *Biochem. Biophys. Res. Commun.*, **297**, 396–400.
24. Zhang,R. and Zhang,C.T. (2003) Multiple replication origins of the archaeon *Halobacterium* species NRC-1. *Biochem. Biophys. Res. Commun.*, **302**, 728–734.
25. Berquist,B.R. and DasSarma,S. (2003) An archaeal chromosomal autonomously replicating sequence element from an extreme halophile, *Halobacterium* sp. strain NRC-1. *J. Bacteriol.*, **185**, 5959–5966.
26. Coker,J.A., DasSarma,P., Capes,M., Wallace,T., McGarrity,K., Gessler,R., Liu,J., Xiang,H., Tatusov,R. and Berquist,B.R. (2009) Multiple replication origins of *Halobacterium* sp. strain NRC-1: properties of the conserved *orc7*-dependent *oriC1*. *J. Bacteriol.*, **191**, 5253–5261.

27. Robinson,N.P., Dionne,I., Lundgren,M., Marsh,V.L., Bernander,R. and Bell,S.D. (2004) Identification of two origins of replication in the single chromosome of the archaeon *Sulfolobus solfataricus*. *Cell*, **116**, 25–38.
28. Lundgren,M., Andersson,A., Chen,L., Nilsson,P. and Bernander,R. (2004) Three replication origins in *Sulfolobus* species: synchronous initiation of chromosome replication and asynchronous termination. *Proc. Natl Acad. Sci. USA*, **101**, 7046–7051.
29. Myllykallio,H., Lopez,P., Lopez-Garcia,P., Heilig,R., Saurin,W., Zivanovic,Y., Philippe,H. and Forterre,P. (2000) Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science*, **288**, 2212–2215.
30. Matsunaga,F., Norais,C., Forterre,P. and Myllykallio,H. (2003) Identification of short 'eukaryotic' Okazaki fragments synthesized from a prokaryotic replication origin. *EMBO Rep.*, **4**, 154–158.
31. Norais,C., Hawkins,M., Hartman,A.L., Eisen,J.A., Myllykallio,H. and Allers,T. (2007) Genetic and physical mapping of DNA replication origins in *Haloferax volcanii*. *PLoS Genet.*, **3**, e77.
32. Robinson,N.P. and Bell,S.D. (2007) Extrachromosomal element capture and the evolution of multiple replication origins in archaeal chromosomes. *Proc. Natl Acad. Sci. USA*, **104**, 5806–5811.
33. Majernik,A.I. and Chong,J.P. (2008) A conserved mechanism for replication origin recognition and binding in archaea. *Biochem. J.*, **409**, 511–518.
34. Pelve,E.A., Lindås,A.C., Knöppel,A., Mira,A. and Bernander,R. (2012) Four chromosome replication origins in the archaeon *Pyrobaculum calidifontis*. *Mol. Microbiol.*, **85**, 986–995.
35. Lopez,P., Philippe,H., Myllykallio,H. and Forterre,P. (1999) Identification of putative chromosomal origins of replication in Archaea. *Mol. Microbiol.*, **32**, 883–886.
36. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.